AFRL-IF-RS-TR-2006-88
Final Technical Report
March 2006

MONITORING BUSINESS ACTIVITY

New York University

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

STINFO FINAL REPORT

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2006-88 has been reviewed and is approved for publication.

APPROVED: /s/

DEBORAH A. CERINO
Project Engineer

FOR THE DIRECTOR: /s/

JOSEPH CAMERA
Chief, Information & Intelligence Exploitation Division
Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

1. AGENCY USE ONLY (Leave blank)	**2. REPORT DATE** MARCH 2006	**3. REPORT TYPE AND DATES COVERED** Final Sep 01 – Oct 05

4. TITLE AND SUBTITLE
MONITORING BUSINESS ACTIVITY

5. FUNDING NUMBERS
C - F30602-01-2-0585
PE - 31011G
PR - EELD
TA - 01
WU - 09

6. AUTHOR(S)
Foster Provost, Sofus Macskassy

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
New York University
70 Washington Square South
New York New York 10012-0091

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory/IFED
525 Brooks Road
Rome New York 13441-4505

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-2006-88

11. SUPPLEMENTARY NOTES

AFRL Project Engineer: Deborah A. Cerino/IFED/(315) 330-1445/ Deborah.Cerino@rl.af.mil

12a. DISTRIBUTION / AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

Under this project, we studied and developed technologies to “score” entities to build models that will produce an estimate of the likelihood that an entity exhibits some characteristic. For example, a social network may include malicious individuals. Suspicion scoring assigns a numeric value to each entity in the network, representing the estimated likelihood that the entity is malicious. We have focused on scoring entities that are interconnected in some sort of a network (e.g., a social network) and on techniques for building and using scoring models when important information is unknown, but may be acquired at a cost. This project has also produced 20 published technical papers.

14. SUBJECT TERMS
Networked data, relational neighbor algorithms, scoring networked entities

15. NUMBER OF PAGES
21

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Table of Contents

Executive Summary 1

1 Introduction 2

2 Scoring in networked data: NetKit, Relational Neighbors and suspicion scoring 4

2.1 Classification and scoring in networked data: NetKit and Relational Neighbors 4

2.2 Suspicion scoring on terrorist-world data, with targeted data acquisition 5

2.3 Additional NetKit information (not discussed in publications/references) 7

2.4 NetKit Transition Record 7

2.5 Some NetKit technical details not appearing in the published papers 7

2.6 Associated references and corresponding citations 8

3 ACORA and modeling/scoring with identifier attributes 10

3.1 ACORA overview 10

3.2 Transition note 11

3.3 Associated appendices and corresponding citations 11

4 Active acquisition of information for modeling and scoring 11

4.1 Active sampling for class probability estimation and ranking 12

4.2 Taking the decision-making task into account 12

4.3 Active feature-value acquisition 12

4.4 Inference with missing features 13

5 Other work 13

5.1 The 2003 KDD Cup 13

5.2 Scoring ability versus the amount of data available 14

5.3 Adding numbers to text classification 15

6 Miscellaneous 15

6.1 Performance Evaluation 15

6.2 Relation to Air Force Program Concept of Operations 15

7 Acknowledgement & Disclaimer 16

8 List of references and corresponding citations 16

9 Additional references 17

Executive Summary

Under this project we studied and developed technologies to "score" entities—to build models that will produce an estimate of the likelihood that an entity exhibits some characteristic. For example, a social network may include malicious individuals. Suspicion scoring assigns a numeric value to each entity in the network, representing the estimated likelihood that the entity is malicious. We have focused on scoring entities that are interconnected in some sort of a network (e.g., a social network) and on techniques for building and using scoring models when important information is unknown, but may be acquired at a cost. This project has produced twenty published technical papers; given the volume of results produced, the main body of this report provides background, motivation, and high-level descriptions of techniques and results. The technical details are presented in the references.

This project has built an integrated toolkit, called Netkit, of methods for scoring networked entities, relaxing the standard assumption that entities to be scored are independent (technically, i.i.d. or independent and identically distributed). NetKit has been applied to various benchmark networked data sets, showing that simple methods alone can produce remarkably good scores. Additional development and experimentation was conducted with NetKit's Relational Neighbor (RN) algorithms, which combine a form of guilt-by-association with collective inferencing—in which the entire network is scored simultaneously, so that *scores* of related entities can affect each other. The RN algorithms were applied to the terrorist-world simulation data produced under another project within this Program by Global InfoTek Inc. Early on in the project, the RN algorithms simply found all the suspicious entities. This led to modifications in the development of the simulator. At the end of the project, RN had varying success on the simulated data, depending on the level of noise and observability. As mentioned above, one important aspect of this project is that often information must be acquired at a cost. The results also show the effectiveness of an information-gathering policy that expends resources to acquire more information about those individuals currently deemed to be the most suspicious.

The Automated Construction of Relational Attributes (ACORA) system addresses a particular characteristic of building and using scoring models with networked data, and other relational data. Often the specific identities of entities (individuals or objects) can be very important for scoring, rather than just the characteristics of the entities. For example, to have met with a specific individual may be telling. However, traditional modeling, as well as prior relational techniques, would avoid dealing with high-dimensional categorical variables (like names) due to the explosion of the size of the hypothesis space, and the concomitant danger of overfitting. Under this project we introduced techniques for automatically constructing attributes from high-dimensional categorical attributes, and show that they can consistently and sometimes dramatically improve modeling and scoring.

We also have produced a collection of techniques and results focused on the problem of how to utilize information-gathering resources most cost-effectively, when building and using classification/scoring models. Specifically, we developed techniques for intelligently and judiciously acquiring data/information, in order to improve modeling and scoring when important information is unknown, but can be acquired at a cost. Most of these techniques involve interleaving modeling with "actively" acquiring data, in a variety of usage scenarios.

Finally, we present a collection of related results, such as those that won second and third places in the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining competition, on modeling relational data, as well as some others.

1 Introduction

We have studied and developed technologies for "scoring" entities—building models that will produce an estimate of the likelihood that an entity exhibits some characteristic. *Suspicion scoring* assigns a numeric value to each entity in the network, representing the estimated likelihood that the entity is malicious. For example, for detecting telecommunications fraud it is useful to score accounts in terms of their likelihood of exhibiting fraudulent activity [Fawcett & Provost, 1997]. This project has focused on scoring entities that are interconnected in a network (e.g., a social network) and on techniques for building and using scoring models when important information must be acquired at a cost.

More generally, we estimate probability distributions over the categorical values of entities' attributes. The attribute to be estimated is called the *target* attribute, and often represents "class" membership, such as whether an account belongs to the class of "defrauded accounts" (i.e., the value of the binary attribute `fraud?` is `true`) or whether an academic paper belongs to the class "neural network papers" (i.e., the value of the attribute `paper_topic` is `neural_networks`). In this report, we will use the term "classification" as shorthand for this type of scoring.[1] The general goal is to have a model that will enable fast, accurate scoring of entities (individuals or objects) that are interconnected in some sort of a network (e.g., a social network). If there are enough data for which the value of the target variable (the *label*) is known in advance, then the goal is to *learn* scoring models—that is, to build models automatically from these labeled *training* data.

When data are interconnected, opportunities arise that are not available for standard modeling. Standard statistical and machine-learning techniques induce a model—in our case a classification or a class-probability estimation model—that maps a set of characteristics of an entity to a prediction for the (unknown) value of the target attribute. For example, the address and calling plan and current balance and activity summary may be used to help detect fraud. Let us call this set of characteristics the "local" attributes of the particular entity. In addition to the local attributes, with network data we have the opportunity to take into account the attributes of the entities that are connected to the to-be-classified entity. Therefore, for example, we can take into account the fact that a person has called another account that turned out to be defrauded. We will call the attributes of related entities "relational attributes."

A further opportunity presented by relational data is that *estimates* of the values for unknown relational attributes (including estimates of neighbors' labels) can be used during inference. For example, we may not know the fraud status of any accounts to whom our to-be-classified account has talked, but our model may estimate that one or more is quite suspicious. This may be quite useful in classification. Of course, it introduces a chicken-and-egg problem: my estimate may affect my neighbor's, which in turn may affect mine. Techniques for (quasi-)simultaneously estimating values for all entities are called techniques for "collective inference" [Jensen, et al., 2004]. In the context of suspicion scoring, these techniques can be viewed roughly as propagating suspicion through the network.

A third opportunity presented by relational data, such as networked data, is that specific *identities* of related entities (individuals or objects) can be very important for classification. For example, to have called a specific individual may be telling. However, standard modeling practice avoids dealing with high-dimensional categorical variables (like names) due to the explosion of the size of the hypothesis space, and the concomitant danger of overfitting, as well as representational complications when entities can be connected to an arbitrary number of other entities (who in turn can be connected to an arbitrary number of other entities, and so on). Generally, in relational modeling one must *aggregate* information from multiple, related entities.

We address these issues of learning and inference in networked data in two, loosely connected bodies of work, which are described in the following two sections. First we will describe the NetKit toolkit for

[1] In cases where we want to distinguish the actual prediction of a particular categorical value, we will say "categorical classification."

learning in network data, the Relational Neighbor algorithms, and various experiments performed on benchmark data sets, as well as data from a terrorist-world simulator created (not by us) as part of the overall Air Force umbrella program. In the subsequent section we will describe our work on automatically creating new attributes in the face of relational data with (useful) identifier attributes, including the ACORA system and related experiments.

We also address an orthogonal problem: how to utilize information-gathering resources most cost-effectively, when building classification/scoring models. A characteristic of modeling in many domains is that all the requisite information/data is not available when the modeling process starts. Some data (or other information) may be acquired at a cost. Therefore, given that one does not have infinite resources to acquire data, an important question is: which specific data should be acquired. One general method for addressing this question is "active" modeling: interleaving modeling with targeted data acquisition. For example, traditional active learning techniques produce categorical classification models from whatever labeled data are available, and then use this model to identify one or more "unlabeled" cases for which it would be most worthwhile to invest in acquiring the value of the target variable. The process can iterate indefinitely (e.g., until a budget is exhausted). For scoring, we need similar techniques to best acquire data to improve estimates of the probability of class membership. It is also important to extend these techniques to the acquisition of data other than values of the target variable.

One active data acquisition technique is specifically related to suspicion-scoring in networked data, and will be presented along with the Relational Neighbor technique in the next section. Then, in Section 4, we describe new techniques for various data acquisition scenarios that had not previously been addressed, and corresponding experiments.

Finally, we will summarize various other related results that do not fit cleanly into one of these three main areas of focus. For example, we were encouraged by our Program Manager to compete in the KDD Cup competition in 2003, which happened to involve networked data (a large citation network of physics papers). We earned second- and third-place finishes. The second-place spot was for the task of predicting the number of citations an academic paper would receive in the next quarter, where we constructed features for an ordered probit (i.e., probability unit) model. Our third-place finish was for the "Open Task" (make up an interesting problem and then solve it), where we showed how paper authors can be identified quite well using only the citations they make ("The Myth of the Double-blind Review?"). Separately, we created network-based industry classification models—which deserve further attention, but became peripheral to our efforts on this project. Also, we performed a large-scale, comprehensive study of scoring models based on (bagged) tree induction and logistic regression, in part because it is important to understand the relative abilities of standard scoring algorithms, if they are going to become the basis for more complex, network-based scoring. This study showed some remarkable relationships between characteristics of the data and classification/scoring performance.

This project has produced twenty published technical papers. Given the volume of results produced by the project, the sections that follow in the main body of the report provide background, motivation, and high-level descriptions of techniques and results. The technical details are presented in the references.

2 Scoring in networked data: NetKit, Relational Neighbors and suspicion scoring

We first will motivate and describe the network toolkit NetKit, including the Relational Neighbor algorithms. This work is described in detail in Reference A, which is a paper accepted for publication (with revisions) to the *Journal of Machine Learning Research (JMLR)*.[2] Then we will describe the application of NetKit's Relational Neighbor algorithm to data from the Program's terrorist-world simulator, including a method we introduced for targeted acquisition of costly secondary data. This work was published in the 2005 International Conference on Intelligence Analysis.

From the perspective of intelligence analysis, the goal of this work is to develop methods and tools that will help to reduce (substantially) the necessary manual processing of data, and at the same time to increase the accuracy of the suspicion scores that are produced—meaning that the suspicion scores better correlate with the likelihood of engaging in malicious activity.

Suspicion scoring is a key ingredient of an Intelligence Analysis Knowledge Base (IAKB), which has been identified as one of the top-10 needs for analysts [Badalamente & Greitzer, IA-2005]. Within the overall Air Force umbrella Program, our suspicion scores were taken as input and used by other contractors (viz., by 21st Century and by Alphatech).

2.1 Classification and scoring in networked data: NetKit and Relational Neighbors

NetKit-SRL, or NetKit for short, is a modular toolkit for classification in networked data.[3] We describe here and in Reference A the toolkit and a case-study of its application to a collection of networked data sets used in prior machine learning research.

Networked data are relational data where entities are interconnected, and we consider the case where entities whose labels are to be estimated are (sometimes) linked to entities for which the label is known. For example, a wireless phone account for which we would like to estimate the fraud status may have called another account for which we know the fraud status. NetKit is based on a three-component framework, comprising a local classifier, a relational classifier, and a collective inference procedure. Various existing relational learning algorithms can be instantiated with appropriate choices for these three components and new relational learning algorithms can be composed by new combinations of components.

NetKit is interesting for several reasons. First, it encompasses several currently available systems, which are realized by choosing particular instantiations for the different components. This allows us to compare and contrast the different systems on equal footing. Perhaps more importantly, the modularity of the toolkit broadens the design space of possible systems beyond those that have appeared in prior published work, either by mixing and matching the components of the prior systems, or by introducing new alternatives for components. We also introduce novel "Relational Neighbor" algorithms, which we describe next and more fully in the reference. Finally, NetKit's modularity not only allows for direct comparison of various models, but also for comparison of isolated components, as we show.

The Relational Neighbor algorithms are a relational analogy to the long-popular Nearest Neighbor algorithms for non-parametric statistical inference. The idea is to base the classification of an entity on the (known or inferred) classifications of the entities to which it is connected. Technically, the relational neighbor algorithm applies an aggregation operator to the classes of the neighbors. A simple aggregation would be to take the majority or plurality class; a slightly more complex operator could aggregate with a weighted voting. For scoring, the wvRN algorithm uses a normalized sum of the weights of the neighbors exhibiting the class in question, based on fixed weights, and the cdRN algorithm tries to learn how to combine the weighted classes (see the reference for details). In order to be applicable to cases where some

[2] Currently ranked by ISI as the 2nd highest impact journal in Computer Science.

[3] NetKit-SRL, or NetKit for short, is written in Java 1.5 and is available for research purposes as open source.

(or all) neighbors' labels are unknown, the RN algorithms are combined with any of the collective inference methods. The case study suggests that RN excels when combined with relaxation labeling. In the sequel, unless otherwise indicated, RN corresponds to wvRN with relaxation labeling.

To illustrate NetKit's benefits, and to evaluate the various algorithms, we have used the toolkit to conduct an in-depth case study of within-network classification. We compare various techniques on twelve benchmark data sets from four domains used in prior machine learning research. Beyond illustrating the value of the toolkit, the case study makes several contributions. The case study demonstrates how the toolkit facilitates comparison of different learning methods (which so far has been lacking in machine learning research). It also shows how the modular framework allows analysis of subcomponents, to assess which, whether, and when particular components contribute to superior performance.

The case study focuses on the simple but important special case of univariate network classification, for which the only information available is the structure of class linkage in the network (i.e., only links and some class labels are available). To our knowledge, no work previously has evaluated systematically the power of class-linkage alone for classification (in benchmark data sets). The results demonstrate clearly that simple network-classification models perform remarkably well—well enough that they should be used regularly as baseline classifiers for studies of relational learning for networked data.

The results also show that there are a small number of component combinations that excel, and that different components are preferable in different situations. For example, there is a clear phase-shift between when very simple methods perform better and when more complex components perform better, based on how sparse existing target-variable information is in the network. Specifically, we tested for sensitivity to initial conditions, varying the amount of initially labeled data from 10% to 90% of a graph, where the task was to label the remaining nodes in the graph (those not initially labeled.) We found that two simple methods, the simplest being wvRN, performed exceedingly well when paired with the relaxation-labeling method for collective inference. This combination was significantly better than any other combination when less than half of the graph was initially labeled. When more than half of the graph was labeled initially, then a logistic regression-based network classification method performed the best, regardless of the collective inference method with which it was paired.

Two crucial characteristics are needed in a network in order for the simplest within-network classification method to work in the univariate case: unlabeled nodes must be (in)directly connected to labeled nodes and the network must exhibit some level of homophily—i.e., they are more likely to be connected to nodes of a similar class than not (terrorists are more likely to connect to other terrorists than are non-terrorists). Interestingly, we have observed both of these characteristics across all our benchmark data sets, ranging from social networks to films to research papers to publicly traded companies.

One question that arises from this study is: if various different ways are available to link entities, and one wants to run simple methods on a homogeneous network (with only one sort of link), how should the links be chosen? We demonstrate analogues to traditional feature selection that select good (often the best) link definitions.

Reference A provides (much) detail, describing the problem of network learning formally, introducing the modular framework in technical detail, reviewing prior work, and describing NetKit technically. Then it describes the case study, including the experimental methodology, data used, toolkit components used, and the results and analysis of the comparative study. The reference concludes with discussions of limitations and conclusions. We also include, as additional appendices, various predecessor papers, including position papers and papers that led up to the development of NetKit. (See Section 2.6).

2.2 Suspicion scoring on terrorist-world data, with targeted data acquisition

In the context of the EAGLE program, we used NetKit to study *suspicion scoring*: ranking individuals by their estimated likelihood of being malicious. In particular, we addressed suspicion scoring in networks of people, linked by communications, meetings, or other associations (e.g., being in the same vicinity at the same time). Our system makes use of the simple-yet-ubiquitous principle of homophily [Blau 1977;

McPherson et al. 2001]; social research has shown repeatedly that a person is more likely to associate with people who share similar interests or characteristics. Homophily is the basis of a simple guilt-by-association algorithm: estimate suspicion level by counting malicious associates.

Previously, for fraud detection, suspicion scoring based on networked data has been used successfully, although typically in an ad hoc manner. The "dialed digits" monitors discussed by Fawcett and Provost give an account a high score if it calls the same numbers called by known fraudulent accounts [Fawcett and Provost 1997]; the "communities of interest" of Cortes et al. explicitly represent the network neighborhoods around telephone accounts as a basis for suspicion scoring [Cortes et al. 2001]. We extend such methods by propagating suspicion through the association network, and conducting suspicion-based acquisition of additional data.

One problem with using the simple homophily-based guilt-by-association algorithm in large networks is that *few people may be known to be malicious*. Often none of an individual's associates are known to be either malicious or benign. However, if the association graph is well connected, then following linkages of associations is likely eventually to lead to at least one individual who is known or strongly suspected to be malicious. Based on this idea, we overcome the problem of sparse knowledge by propagating suspicion scores through the association network until all suspicion scores stabilize. In particular, we use an adaptation of the relaxation labeling method shown to yield good performance for hypertext classification by Chakrabarti et al. [1998].

Relaxation labeling works well if the association graph is well-connected. For intelligence data, one must consider the difference between the *true* association network, and the network of *known* associations. *The true association network may be known only partially*. We address this via suspicion-based data acquisition, using current suspicion scores to acquire data on additional connections in order to improve the suspicion propagation. In a realistic setting, acquiring association links (involving subpoenas for transaction records, surveillance, interviews, phone taps, etc.) is costly in terms of money, resources, legal issues, and public perception. We attempt to minimize costs by acquiring such "secondary data" only for entities with the highest estimated suspiciousness. This heuristic works well in the data studied.

We demonstrated the method on a suite of data sets generated by the Air Force program's terrorist-world simulator. The data sets consist of thousands of people and some known links between them. We show that the system ranks truly malicious individuals highly, even if only relatively few are known to be malicious ex ante (i.e., beforehand). In models where there is uncertainty that is resolved during the course of events, the ex antes values (e.g. of expected gain) are those that are calculated in advance of the resolution of uncertainty. When used as a tool for identifying promising data-gathering opportunities, the system focuses on gathering more information about the most suspicious people and thereby increases the density of linkage in appropriate parts of the network. We assess performance under conditions of noisy prior knowledge (score quality varies by data set under moderate noise), and whether augmenting the network with prior scores based on profiling information improves the scoring (it doesn't). Although the level of performance reported would not support direct action on all data sets, it does recommend the consideration of network-scoring techniques as a new source of evidence in decision making. For example, the system can operate on networks far larger and more complex than could be processed by a human analyst.

We conducted a preliminary investigation into augmenting the scoring with other uncertain-but-better-than-random knowledge (as from a profiling technique). The priors had little-to-no effect due to the dominance of the scores propagated from the static labels. This is a problem that can have an impact on many collective inference techniques. An important open question is how one should combine relational and local information properly such that one does not dominate the other.

The techniques and results are described in detail in Reference B, which is a paper that appeared at the 2005 *International Conference on Intelligence Analysis*.

2.3 Additional NetKit information (not discussed in publications/references)

We have developed NetKit as a publicly available relational learning toolkit and already have had downloads from around the world. NetKit also is currently undergoing evaluation by intelligence agencies as a potential tool to help in counterterrorism and related domains.

- We released the second version of our Network Learning Toolkit for Statistical Relational Learning (NetKit-SRL). NetKit is written in Java 1.5 and is available as open-source. The toolkit is mature enough to be used by the general public, although we have plans for many enhancements and expansions. The source code is available for download from coPI Sofus Macskassy (sofmac@fetch.com).

- We presented the toolkit and its applications at the Statistical Relational Learning workshop at Dagstuhl in Germany in February 2005, at Google R&D in February 2005, at Brigham Young University in March 2005, at Notre Dame University in April 2005, at the Snowbird Machine Learning workshop in Utah in April, 2005, and at the NAACSOS conference in June 2005. We have received positive feedback and requests for enhancements from leading researchers in Machine Learning.

- NetKit already has been downloaded by individuals and/or groups in academia both in the U.S. as well as in Europe (we have had dialogues with persons both in the U.K. and in Germany). We do not have complete access to the download logs from our serving website, so we cannot give accurate numbers on the number of downloads or their origin.

2.4 NetKit Transition Record

Netkit has been through several stages of evaluation at the Intelligence Community's Research and Development Experimental Collaboration (RDEC) facility:

- Initial prototype delivered December 2004.
- NetKit was promoted to DPP on April 2005.
- RDEC received the new version May 2005.
- On Nov. 17, 2005 we received the following message, indicating a desire to move NetKit to the EP:

```
From: Hydock Thomas [mailto:hydock_thomas@bah.com]
Sent: Thursday, November 17, 2005 8:58 AM
To: Sofus Macskassy
Subject: More Questions

Sofus,

I demo'd NetKit yesterday to some of our RDEC analysts to get some
feedback and an idea of how they would want to use the tool.  They are
interested in bringing NetKit into the EP for further experimentation
against some of their classified datasets.

<technical questions omitted>

Thank you.

- T.C.

***********************************
T.C. Hydock
Senior Consultant
Booz Allen Hamilton
hydock_thomas@bah.com
(703) 902-4124
```

- We have followed up on this request; we have no additional status information as of this writing.

2.5 Some NetKit technical details not appearing in the published papers

The current version of NetKit can be used to generate predictive models on networks of homogeneous nodes and heterogeneous edges. For example, web-pages, citation-graphs and social networks, where the nodes in the graph are all of the same type, but the links between nodes can be different semantically (e.g., co-authorship, citation, friendship, links-to, linked-from). Nodes can have attributes.

Specifically, features of this version include:

- Support for multiple attributes on nodes (e.g., hair-color, eye-color, age, etc.).
- Support for heterogeneous edges (e.g., “friend-of”, “seen-together”, etc.).
- A meta-learner which combines multiple models to generate its predictions (for example to combine predictions from a model that uses only profiling and a model that uses only relations).
- Aggregation of neighbor-attributes (for example, what is the ratio of persons through the “friends-of” relation that have blue eyes).
- Re-vamped I/O resulting in 1-2 orders of magnitude speedup in reading in graphs (significant for very large graphs consisting of 100K’s of nodes and millions of edges).

2.6 Associated references and corresponding citations

- **Reference A:** Macskassy, S. and F. Provost (2004). “Classification in Networked Data: A toolkit and a univariate case study.” Accepted to *Journal of Machine Learning Research*, pending revisions. CeDER Working Paper #CeDER-04-08, Stern School of Business, New York University, NY, NY 10012.
    - *The main technical paper describing NetKit, the philosophy, toolkit, and results.*
- **Reference B:** Macskassy, S. and F. Provost (2005). “Suspicion scoring based on guilt-by-association, collective inference, and focused data access.” Proceedings of the International Conference on Intelligence Analysis.
    - *This paper applies NetKit’s Relational Neighbor classifier, in combination with relaxation labeling and a method for focused data access, to conduct suspicion scoring to rank possible terrorists in the EAGLE Challenge Problem data.*
- **Reference C:** Macskassy, S. and F. Provost (2005). “NetKit-SRL: A Toolkit for Network Learning and Inference.” Proceedings of the Proceedings of the North American Association for Computational Social and Organizational Science (NAACSOS) Conference 2005.
    - *A short (invited) paper introducing NeKit to the NAACSOS audience.*
- **Reference D:** Macskassy, S. and F. Provost (2005). “Suspicion scoring based on guilt-by-association, collective inference, and focused data access.” Proceedings of the North American Association for Computational Social and Organizational Science (NAACSOS) Conference 2005.
    - *A follow-up study to the Reference B paper (that somehow ended up with the same title, unfortunately). It addresses some methodological shortcomings and presents new, improved results.*
- **Reference E:** Bernstein, A., S. Clearwater, S. Hill, C. Perlich, and F. Provost (2002). “Discovering Knowledge from Relational Data Extracted from Business News.” In Proceedings of the KDD-2002 Workshop on Multi-Relational Data Mining.
    - *Early results from this project, not mentioned above. We had proposed (prior to 9/11) to use the business news as a domain that mirrored (technically) many of the characteristics of the problems of interest to the sponsor.*
    - *Using a combination of information extraction, network analysis, and statistical techniques, we show that relationally interlinked patterns distributed over multiple documents indeed can be extracted, and (specifically) that knowledge about companies’ interrelationships can be discovered. We evaluate the extracted relationships in several ways: we give a broad visualization of related companies, showing intuitive industry clusters; we use network analysis to ask who are the central players, and finally, we show that the extracted interrelationships can be used for important tasks, such as classifying companies by industry membership.*
- **Reference F:** Bernstein, A., S. Clearwater, and F. Provost (2003). "The Relational Vector-space Model and Industry Classification." In *Proceedings of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data*.

  - *A followup to the prior study, not discussed above, and the origin of the Relational Neighbor classifier. This paper introduces a relational vector-space (VS) model (in analogy to the VS model used in information retrieval) that abstracts the linked structure, representing entities by vectors of weights. Given labeled data as background knowledge/training data, classification procedures can be defined for this model, including a straightforward, "direct" model using weighted adjacency vectors. Using a large set of tasks from the domain of company affiliation identification, we demonstrate that such classification procedures can be effective. We then examine the method in more detail, showing that as expected the classification performance correlates with the relational autocorrelation of the data set. We then turn the tables and use the relational VS scores as a way to analyze/visualize the relational autocorrelation present in a complex linked structure. The main contribution of the paper is to introduce the relational VS model as a potentially useful addition to the toolkit for relational data mining. It could provide useful constructed features for domains with low to moderate relational autocorrelation; it may be effective by itself for domains with high levels of relational autocorrelation, and it provides a useful abstraction for analyzing the properties of linked data.*
- **Reference G:** Provost, F., C. Perlich, and S. Macskassy (2003). "Relational Learning Problems and Simple Models." In *Proceedings of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data*.
  - *This paper discusses some technical details associated with relational modeling, and argues for the broad use of relational neighbor classifiers in relational learning studies.*
  - *One aspect that has not received much attention previously is that in network data, "training" data are intertwined with those data for which predictions must be made. One implication is that the same data that are used for training, can (often) be used as "background knowledge" during prediction. This leads into the ACORA work, which is presented next, in Section 3.*
- **Reference H:** Macskassy, S. and F. Provost (2003). "A Simple Relational Classifier." In Proceedings of the KDD-2003 Workshop on Multi-Relational Data Mining (MRDM-2003).
  - *A first paper on the Relational Neighbor classifier, per se, including results showing that it can perform remarkably well.*

3 ACORA and modeling/scoring with identifier attributes

In complex domains, predictive modeling often is faced with important relationships between entities. For example, suspicious people may make phone calls to the same numbers as other suspicious people; customers engage in transactions which involve products. Extending traditional "propositional" modeling approaches to account for such relationships introduces a variety of opportunities and challenges. The focus of this work is one such challenge—the integration of information from one-to-many and many-to-many relationships: a person may have called many numbers; a customer may have purchased many products. ACORA (Automated Construction of Relational Attributes) is a system that converts a relational domain into a feature-vector representation using aggregation to construct attributes automatically.

Identifier attributes—very high-dimensional categorical attributes such as people's names or particular product ids—rarely are incorporated in statistical modeling. However, they can play an important role in relational modeling: it may be informative to have communicated with a particular set of people or to have purchased a particular set of products. A key limitation of existing relational modeling techniques is how they *aggregate* bags (multisets) of values from related entities. The main contribution of our work in this area is the introduction of aggregation operators that capture more information about the value distributions, by storing meta-data about value distributions and referencing this meta-data when aggregating—for example by computing class-conditional distributional distances. Such aggregations are particularly important for aggregating values from high-dimensional categorical attributes, for which the simple aggregates provide little information.

3.1 ACORA overview

More technically, the aggregation operators used by existing relational modeling approaches typically are simple summaries of the distributions of features of related entities, e.g., MEAN, MODE, SUM, or COUNT. These operators may be adequate for some features, but fail miserably for others. In particular, if the bag consists of values from high-dimensional categorical attributes, simple aggregates provide little information. Object identifiers are one instance of high-dimensional categorical attributes, and they are abundant in relational domains since they are necessary to express the relationships between objects. Traditional propositional modeling rarely incorporates object identifiers, because they typically hinder generalization (for example by creating "lookup tables"). However, the identities of related entities can play an important role in relational modeling: it may be informative to have communicated with a specific set of people or to have purchased a specific set of products. For example, [Fawcett & Provost, 1997] show that incorporating particular called-numbers, location identifiers, etc., can be quite useful for fraud detection.

Relational learning methods address the need for more automation and support of modeling in such domains, including the ability to explore information about the many-to-many relationship between customers and products. If the modeling objective is to estimate the likelihood of responding to an offer for a particular book, it may be valuable to incorporate the specific books previously bought by the customer, as captured by their ISBNs. The MODE clearly is not suitable to aggregate a bag of ISBNs, since typically books are bought only once by a particular customer. In addition, this MODE feature would have an extremely large number of possible values, perhaps far exceeding the number of training examples.

We introduce novel aggregators that allow learning techniques to capture information from identifiers such as ISBNs. This ability is based on (1) the implicit reduction of the dimensionality by making (restrictive) assumptions about the number of distributions from which the values were generated, and (2) the use of *distances* to class-conditional, distributional meta-data. Such distances reduce the dimensionality of the model estimation problem while maintaining discriminability among instances, and they focus explicitly on discriminative information.

The contributions of this work include:

- An analysis of principles for developing new aggregation operators.

- The development of a novel method for relational feature construction, based on the foregoing analysis, which includes novel aggregation operators. To our knowledge, this is the first relational aggregation approach that can be applied generally to categorical attributes with high cardinality.

- A theoretical justification of the approach that draws an analogy to the statistical distinction between random- and fixed-effect modeling, and identifies typical aggregation assumptions that limit the expressive power of relational models.

- A theoretical conjecture that the aggregation of identifier attributes can implicitly support the learning of models from *unobserved* object properties.

- An extensive empirical study demonstrating that the novel aggregators indeed can improve predictive modeling in domains with important high-dimensional categorical attributes, including a sensitivity analysis of major domain properties.

Reference I (the *Machine Learning* journal paper) presents the technical details supporting all these contributions. In the first half of the paper we provide general guidelines for designing aggregation operators, introduce the new aggregators in the context of the relational learning system ACORA, and provide theoretical justification. We also conjecture special properties of identifier attributes, e.g., they proxy for unobserved attributes and for information deeper in the relationship network. In the second half of the paper we provide extensive empirical evidence that the distribution-based aggregators indeed do facilitate modeling with high-dimensional categorical attributes, and in support of the theoretical conjectures.

3.2 Transition note

The basic aggregation operators used in the ACORA work have been incorporated in the Proximity system for exploring relational data, developed and supported by Prof. David Jensen's laboratory at the University of Massachusetts at Amherst.

3.3 Associated appendices and corresponding citations

- **Reference I:** Perlich, C. and F. Provost (2006). "Distribution-based Aggregation for Relational Learning from Identifier Attributes." *Machine Learning* 62(1/2).
    - *This is the main technical paper describing the ACORA project, theory, methods, and results.*
- **Reference J:** Perlich, C. and F. Provost (2003). "Aggregation and Concept Complexity in Relational Learning." In *Proceedings of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data.*
    - *A position paper related to the ACORA project, arguing that aggregation methods should be a central focus in relational learning and inference.*
- **Reference K:** Perlich, C., and F. Provost (2003). "Aggregation-based Feature Invention and Relational Concept Classes." In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD-2003).
    - This paper discusses different classes of aggregation, presenting them in a hierarchy of increasing complexity. It also introduces the earliest ACORA techniques and results.

4 Active acquisition of information for modeling and scoring

In real-world modeling and inference tasks, often it is necessary to acquire additional information in a focused, active, cost-sensitive manner. We saw this for the suspicion scoring, discussed above in Section 2.2. Relatively speaking, there is not much research on the topic. One of the complications to doing research on active modeling is that there are many different types of information that one might be able to acquire at a cost. In addition to the technique we introduced above for suspicion scoring, we also have developed and evaluated active modeling techniques in various other scenarios. We describe several studies in the following sections. In the final study, we address the (surprisingly open) question of how best to operate when some specific information used by a learned model (viz., the value of a variable used by the model) is missing, at the time when one wants to use the model for inference.

Generally, environments where costs of taking actions (and making mistakes) have to be taken into account are called *cost-sensitive*. Specifically, the term "cost-sensitive learning" usually is used to refer to learning in a particular cost-sensitive environment: a (categorical) classification task where there are costs to making different misclassification errors. In the first work we discuss below (in Section 4.1), we use an extension of this restricted but common setting. Then we address other settings, important for real-world tasks but which have received little or no prior work.

In each section that follows, we give a high-level description of one project, followed by a reference that covers the topic in detail.

4.1 Active sampling for class probability estimation and ranking

In many cost-sensitive environments class probability estimates are used by decision makers to evaluate the expected utility from a set of alternatives. Supervised learning can be used to build class probability estimates; however, it often is very costly to obtain training data with class labels. Active learning acquires data incrementally, at each phase identifying especially useful additional data for labeling, and can be used to economize on examples needed for learning. We outline the critical features of an active learner and present a sampling-based active learning method for estimating class probabilities and class-based rankings. BOOTSTRAP-LV identifies particularly informative new data for learning based on the variance in probability estimates, and uses weighted sampling to account for a potential example's informative value for the rest of the input space. We show empirically that the method reduces the number of data items that must be obtained and labeled, across a wide variety of domains. We investigate the contribution of the components of the algorithm and show that each provides valuable information to help identify informative examples. We also compare BOOTSTRAP-LV with UNCERTAINTY SAMPLING, an existing active learning method designed to maximize classification accuracy. The results show that BOOTSTRAP-LV uses fewer examples to exhibit a certain estimation accuracy and provide insights to the behavior of the algorithms. Finally, we experiment with another new active sampling algorithm drawing from both UNCERTAINTY SAMPLING and BOOTSTRAP-LV and show that it is significantly more competitive with BOOTSTRAP-LV compared to UNCERTAINTY SAMPLING. The analysis suggests more general implications for improving existing active sampling algorithms for classification.

- **Reference L:** Saar-Tsechansky, M. and F. Provost. "Active Sampling for Class Probability Estimation and Ranking." *Machine Learning* 54:2 2004, 153-178.

4.2 Taking the decision-making task into account

Predictive models often are used as part of a decision-making process, and costly improvements in model accuracy do not always result in better decisions. This work develops a new approach for active information acquisition that targets decision-making specifically. The method we introduce departs from the traditional active learning paradigm and places emphasis on acquisitions that are more likely to affect decision-making. Empirical evaluations with direct marketing data demonstrate that for a fixed information acquisition cost the method significantly improves the targeting decisions. The method is designed to be generic—not based on a single model or induction algorithm—and we show that it can be applied effectively to various predictive modeling techniques.

- **Reference M:** Saar-Tsechansky, M. and F. Provost. "Active Learning for Decision-making." CeDER Working Paper #CeDER-04-06.

4.3 Active feature-value acquisition

Many induction problems are missing the values for important variables (features). In most applications, these could be acquired at a cost. For building accurate predictive models, acquiring complete information for all instances may be very expensive or unnecessary. Acquiring information for a random subset of variables or cases may not be most effective. *Active feature-value acquisition* tries to reduce the cost of achieving a desired level of model accuracy by identifying instances for which obtaining complete information is most informative.

In related work, we have developed an active, iterative approach that selects values for acquisition based on the current model's accuracy and its confidence in its predictions. Experimental results demonstrate that our approach can induce accurate models using substantially fewer feature-value acquisitions as compared to alternative policies. This approach is presented in Reference N.

We also have developed an alternative approach, which acquires feature values for inducing a classification model based on an estimation of the expected improvement in model accuracy per unit cost. Experimental results demonstrate that this approach consistently reduces the cost of producing a model of a desired accuracy compared to random feature acquisition. This approach is presented in Reference O.

- **Reference N:** Melville, P., M. Saar-Tsechansky, F. Provost, and R. Mooney (2004). "Active Feature-Value Acquisition for Classifier Induction." In *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM-2004).*
- **Reference O:** Melville, P., M. Saar-Tsechansky, F. Provost, R. Mooney (2005). "Economical Active Feature-value Acquisition through Expected Utility Estimation." In *The Fifth International Conference on Data Mining (ICDM-2005),* and appeared in the *KDD-05 Workshop on Utility-based Data Mining*.

4.4 Inference with missing features

When studying the problem of how to utilize information-gathering resources most cost-effectively, an integral subproblem is how to act most effectively in the absence of certain information. For example, if the value of an important variable—one used by the predictive model—is missing, how should one proceed? For example, diagnostic tests may not be available, or a customer age may be missing. Much work has been done to study the effect of different treatments of missing values on model induction, but little work has been done to evaluate and analyze treatments for missing values at prediction time. This paper experimentally compares several different methods—distribution-based imputation, predictive value imputation, and using reduced models—for applying classification trees to instances with missing values. The resulting accuracies are in inverse relation to the popularity of the methods in AI research and practice. Notably, reduced models consistently outperform the other two treatments, sometimes by a large margin; however, they seldom are used. This in part is due to the (perceived) expense of reduced modeling in terms of computation or storage. In light of these results, we then introduce alternative, hybrid approaches that allow users to balance between more accurate but computationally expensive reduced modeling and the other, less accurate but less computationally expensive treatments.

- **Reference P:** Saar-Tsechansky, M. and F. Provost. "Handling Missing Features when Applying Classification Trees." CeDER Working Paper #CeDER-05-19.

5 Other work

5.1 The 2003 KDD Cup

Every year, the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining holds a competition, called the KDD Cup. Top researchers and practitioners, from academia, from government and industrial labs, and from software and consulting firms, compete to solve the task (best). In 2003, the task involved mining network data. In particular, as described by the 2003 KDD Cup organizers:[4]

> Complex networks have emerged as a central theme in data mining applications, appearing in domains that range from communication networks and the Web, to biological interaction networks, to social networks and homeland security. At the same time, the difficulty in obtaining complete and accurate representations of large networks has been an obstacle to research in this area.

[4] http://www.cs.cornell.edu/projects/kddcup/

> This KDD Cup is based on a very large archive of research papers that provides an unusually comprehensive snapshot of a particular social network in action; in addition to the full text of research papers, it includes both explicit citation structure and (partial) data on the downloading of papers by users. It provides a framework for testing general network and usage mining techniques…
>
> The e-print arXiv, initiated in Aug 1991, has become the primary mode of research communication in multiple fields of physics, and some related disciplines. It currently contains over 225,000 full text articles and is growing at a rate of 40,000 new submissions per year. It provides nearly comprehensive coverage of large areas of physics, and serves as an on-line seminar system for those areas. It serves 10 million requests per month, including tens of thousands of search queries per day. Its collections are a unique resource for algorithmic experiments and model building. Usage data has been collected since 1991, including Web usage logs beginning in 1993. On average, the full text of each paper was downloaded over 300 times since 1996, and some were downloaded tens of thousands of times.
>
> The Stanford Linear Accelerator Center SPIRES-HEP database has been comprehensively cataloguing the High Energy Particle Physics (HEP) literature online since 1974, and indexes more than 500,000 high-energy physics related articles including their full citation tree.

We participated in two tasks:

(1) Predicting the future: Contestants predict (in advance) how many citations each paper will receive during the three months leading up to the KDD 2003 conference.

(2) The open task: Given the large amount of data, contestants can devise their own questions and the most interesting result is the winner.

We won second place in predicting the future, and third place in the open task.

For predicting the future, we focused on data selection and feature construction, to train an ordered probit model. This is described in the (2-page) report available as Reference Q.

For the open task, our entry was entitled "The Myth of the Double Blind Review?" Prior studies have questioned the degree of anonymity of the double-blind review process for scholarly research articles. For example, one study based on a survey of reviewers concluded that authors often could be identified by reviewers using a combination of the author's reference list and the referee's personal background knowledge. We examined how well various automatic matching techniques could identify authors within the competition's very large archive of research papers. Reference R describes the issues surrounding author identification, how these issues motivated our study, and the results we obtained. The best method, based on discriminative self-citations, identified authors correctly 40-45% of the time. One main motivation for double-blind review is to eliminate bias in favor of well-known authors. However, identification accuracy for authors with substantial publication history is even better (60% accuracy for the top-10% most prolific authors, 85% for authors with 100 or more prior papers).

- **Reference Q:** Perlich, C., F. Provost, and S. Macskassy (2003). "Predicting citation rates for physics papers: Constructing features for an ordered probit model." *SIGKDD Explorations* 5(2) 2003, 89-90.
- **Reference R:** Hill, S. and F. Provost (2003). "The Myth of the Double-Blind Review? Author Identification Using Only Citations." *SIGKDD Explorations* 5(2) 2003, 114-119.

5.2 Scoring ability versus the amount of data available

We investigated the classification performance of two standard, off-the-shelf methods for building models for classification and scoring. We conducted a large-scale experimental comparison of logistic regression and tree induction, assessing classification accuracy and the quality of rankings based on class-membership probabilities. We use a learning-curve analysis to examine the relationship of these measures to the size of the training set. The results of the study show several things. (1) Contrary to some prior observations, logistic regression does not generally outperform tree induction. (2) More specifically, and not surprisingly, logistic regression is better for smaller training sets and tree induction for larger data sets. Importantly, this often holds for training sets drawn from the same domain (that is, the learning curves cross), so conclusions about induction-algorithm superiority on a given domain must be based on an analysis of the learning curves. (3) Contrary to conventional wisdom, tree induction is effective at producing probability-based rankings. Finally, (4) the domains on which tree induction and logistic regression are ultimately preferable can be characterized surprisingly well by a simple measure of the separability of signal from noise.

- **Reference S:** Perlich, C., F. Provost, and J. Simonoff (2003). "Tree Induction vs. Logistic Regression: A Learning-curve Analysis." *Journal of Machine Learning Research* 4 (2003) 211-255.

5.3 Adding numbers to text classification

Many real-world problems involve a combination of both text- and numerical-valued features. For example, in email classification, it is possible to use instance representations that consider not only the text of each message, but also numerical-valued features such as the length of the message or the time of day at which it was sent. Text classification methods have thus far not easily incorporated numerical features. One approach for adding numerical features to a text classification problem would create a bag of tokens for each number, such that numbers that are close share many tokens whereas more distant numbers do not. We show, using new benchmark problems in two domains, that this approach is an effective way to learn from both text and numbers and that the bag-of-tokens encoding outperforms a simpler “binning” encoding. Moreover, we show that selecting a best classification method using text-only features and then adding numerical features to the problem (as might happen if numerical features are only later added to a pre-existing text-classification problem) gives performance that rivals a more time-consuming approach of re-evaluating all classification methods using the full set of both text and numerical features.

- **Reference T:** Sofus A. Macskassy, Haym Hirsh (2003). Adding Numbers to Text Classification. In Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM 2003).

6 Miscellaneous

6.1 Performance Evaluation

We have conducted comprehensive, sometimes very large-scale performance evaluations. Our technologies perform well as compared to alternatives—often well enough for publication in the scientific literature. Details are presented in the references.

6.2 Relation to Air Force Program Concept of Operations

Suspicion scoring can be used in many places within a system for generating alerts of potential malicious activity. Of course, it can be used directly to identify individuals for further investigation. It also can be taken as input to various systems that look for malicious groups or activities, providing initial focus for subsequent investigations.

Within the Air Force program, it was rare for the technologies of other contractors to take as input continuous scores (e.g., estimations of probabilities). In many cases, we were asked to provide black-and-white assessments of who is or is not malicious. In our experience with similar applications (e.g., fraud detection), this often simply cannot be done—there is not enough evidence to say with certainty that an actor definitely is malicious or not. The best that can be done is to rank different actors with some sort of score. In the best case, this is a true probability estimate. Which actors subsequently to treat as probably being malicious depends upon the costs of the particular application context. For example, the costs are very different if the scores are used to focus subsequent investigation, as compared to taking a concrete action against the actors (e.g., in the fraud context, shutting down a customer’s account).

Our wvRN-RL scores were used by Alphatech, Inc. in their boundary studies, and by 21st Century on Program challenge problems (as mentioned in a presentation at the 2005 International Conference on Intelligence Analysis).

As discussed in detail above, NetKit has undergone transition through the RDEC process. Also, one component of ACORA has been integrated with UMass-Jensen’s Proximity and is available in their official April-2005 release.

7 Acknowledgement & Disclaimer

The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Department of Defense, the Air Force Research Laboratory, or the U.S. Government.

8 List of references and corresponding citations

- **Reference A:** Macskassy, S. and F. Provost (2004). “Classification in Networked Data: A toolkit and a univariate case study.” Accepted to *Journal of Machine Learning Research*, pending revisions. CeDER Working Paper #CeDER-04-08, Stern School of Business, New York University, NY, NY 10012.
- **Reference B:** Macskassy, S. and F. Provost (2005). “Suspicion scoring based on guilt-by-association, collective inference, and focused data access.” Proceedings of the International Conference on Intelligence Analysis.
- **Reference C:** Macskassy, S. and F. Provost (2005). “NetKit-SRL: A Toolkit for Network Learning and Inference.” Proceedings of the Proceedings of the North American Association for Computational Social and Organizational Science (NAACSOS) Conference 2005.
- **Reference D:** Macskassy, S. and F. Provost (2005). “Suspicion scoring based on guilt-by-association, collective inference, and focused data access.” Proceedings of the North American Association for Computational Social and Organizational Science (NAACSOS) Conference 2005.
- **Reference E:** Bernstein, A., S. Clearwater, S. Hill, C. Perlich, and F. Provost (2002). “Discovering Knowledge from Relational Data Extracted from Business News.” In Proceedings of the KDD-2002 Workshop on Multi-Relational Data Mining.
- **Reference F:** Bernstein, A., S. Clearwater, and F. Provost (2003). "The Relational Vector-space Model and Industry Classification." In *Proceedings of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data*.
- **Reference G:** Provost, F., C. Perlich, and S. Macskassy (2003). "Relational Learning Problems and Simple Models." In *Proceedings of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data*.
- **Reference H:** Macskassy, S. and F. Provost (2003). “A Simple Relational Classifier.” In Proceedings of the KDD-2003 Workshop on Multi-Relational Data Mining (MRDM-2003).
- **Reference I:** Perlich, C. and F. Provost (2006). “Distribution-based Aggregation for Relational Learning from Identifier Attributes.” *Machine Learning* 62(1/2).
- **Reference J:** Perlich, C. and F. Provost (2003) "Aggregation and Concept Complexity in Relational Learning." In *Proceedings of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data.*
- **Reference K:** Perlich, C., and F. Provost (2003). “Aggregation-based Feature Invention and Relational Concept Classes.” In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD-2003).
- **Reference L:** Saar-Tsechansky, M. and F. Provost (2004). "Active Sampling for Class Probability Estimation and Ranking." *Machine Learning* 54:2 2004, 153-178.
- **Reference M:** Saar-Tsechansky, M. and F. Provost. “Active Learning for Decision-making.” CeDER Working Paper #CeDER-04-06.
- **Reference N:** Melville, P., M. Saar-Tsechansky, F. Provost, and R. Mooney (2004). “Active Feature-Value Acquisition for Classifier Induction.” In *Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM-2004).*
- **Reference O:** Melville, P., M. Saar-Tsechansky, F. Provost, R. Mooney (2005). “Economical Active Feature-value Acquisition through Expected Utility Estimation.” In *The Fifth International Conference on Data Mining (ICDM-2005),* and appeared in the *KDD-05 Workshop on Utility-based Data Mining*.

- **Reference P:** Saar-Tsechansky, M. and F. Provost. “Handling Missing Features when Applying Classification Trees.” CeDER Working Paper #CeDER-05-19.
- **Reference Q:** Perlich, C., F. Provost, and S. Macskassy (2003). “Predicting citation rates for physics papers: Constructing features for an ordered probit model.” *SIGKDD Explorations* 5(2) 2003, 89-90.
- **Reference R:** Hill, S. and F. Provost. “The Myth of the Double-Blind Review? Author Identification Using Only Citations.” *SIGKDD Explorations* 5(2) 2003, 114-119.
- **Reference S:** Perlich, C., F. Provost, and J. Simonoff. "Tree Induction vs. Logistic Regression: A Learning-curve Analysis." *Journal of Machine Learning Research* 4 (2003) 211-255.
- **Reference T:** Sofus A. Macskassy, Haym Hirsh (2003). Adding Numbers to Text Classification. In Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM 2003).

9 Additional references

Badalamente, R., & F. Greitzer (2005). Top Ten Needs for Intelligence Analysis Tool Development. Proceedings of the 2005 International Conference on Intelligence Analysis.

Blau, P. (1977) Inequality and Heterogeneity: A Primitive Theory of Social Structure. New York: Free Press, 1977.

Chakrabarti, S., B. Dom, and P. Indyk. (1998) Enhanced Hypertext Categorization Using Hyperlinks. In *ACM SIGMOD International Conference on Management of Data, 1998.*

Cortes, C., D. Pregibon and C. Volinsky (2001) Communities of Interest, The *Fourth International Symposium of Intelligent Data Analysis (IDA 2001),* 2001.

Fawcett, T., and F. Provost (1997) Adaptive Fraud Detection. *Data Mining and Knowledge Discovery, 3,* 291-316, 1997.

Jensen, D., J. Neville and B. Gallagher (2004) Why collective inference improves relational classification. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

McPherson, M., L. Smith-Lovin, and J. M. Cook. (2001) Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology, 27,* 415-444, 2001.